THE UNIVERSITY of York

Discussion Papers in Economics

No. 2000/44

Is More Data Better?

by

Kaushik Mitra

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Is more data better?

Kaushik Mitra*
Department of Economics
University of York
Heslington, York YO10 5DD
United Kingdom
Tel: +44 1904 433750
Fax: +44 1904 433759
email: km15@york.ac.uk

This draft: September 25, 2000

Abstract. Conventional wisdom usually suggests that agents should use all the data they have to make the best possible prediction. In this paper, however, it is shown that agents may sometimes be able to make better predictions by throwing away old data. The optimality criterion agents adopt is the mean squared error criterion.

Journal of Economic Literature Classification Numbers: C13, C22, C53, D83, E32, E37.

***Acknowledgements**: This paper was written when I was a postdoctoral fellow at the Research Unit on Economic Structures and Growth (RUESG), Dept of Economics, University of Helsinki, Finland. Financial support from the Academy of Finland and the Yrjö Jahnsson Foundation is gratefully acknowledged. I am very grateful to Seppo Honkapohja for asking me the question which led to this paper. Apart from benefitting constantly from his comments and suggestions, some of the ideas have also crystallized in the course of collaboration with him on a related project. I am grateful to George Evans for helpful discussions and suggestions. Comments provided by Karim Abadir, Emilio Barucci, John Duffy, Ed Greenberg, Cars Hommes, Manuel Santos, Gerhard Sorger and Gabriel Talmain were useful. The usual disclaimer applies.

1. Introduction

Econometric theory assures us that agents can make good estimations with large sample sizes. It can often be shown that the estimates of model parameters converge asymptotically to the true values. The literature on convergence to rational expectations equilibria under "learning" dynamics also often assures us that agents' expectations converge to rational expectations when they have large data sets. Consequently, conventional wisdom suggests that the use of large amounts of data would be beneficial to agents.

This paper, on the contrary, demonstrates that it may be in the agents' interest to throw away old data to improve their prediction of some relevant variable. Based on a commonly used "optimality" criterion agents may find it profitable to use only "small" amounts of data to predict future prices. Essentially the problem I have in mind is that of a true data generating process given by some Markovian process which is assumed unknown to the economic agent. Agents forecast the next realization of this process by using the sample (arithmetic) mean of a certain fixed number of observations of the process. The justification for using the sample mean are several fold. For one thing, it is an unbiased estimator of the (unknown) population mean. For another thing, the law of large numbers of Markov processes implies that the sample mean is expected to converge to the mean of the (asymptotic) distribution of the true process with large enough data. However, while the use of a large amount of data may be good from the point of view of learning the true population mean, I show that this is not necessarily so if agents are interested solely in forecasting the realization of this process.

I also show that there are a couple of economic example models which neatly fit the framework of the paper. The first model can be interpreted to describe the behavior of a firm producing in a perfectly competitive market. The firm chooses output based on its forecast of the price last period to maximize asymptotic expected profits every period. The firm is interested in the amount of data to use to maximize expected profits. The second model follows the permanent income hypothesis considered by Lucas (1976). The agent here wants to forecast his entire future income stream based on previous data of his income and wants to minimize the expected squared prediction error made in the forecast.

The paper is organized as follows. The basic problem is set out and the question of optimal memory length is studied in Section 2. Section 3 describes some economic models which fit the framework of Section 2. Section 4 examines whether agents might detect some mis-specification in their model by using the concept of "consistent" expectations introduced by Hommes and Sorger (1998). The final section discusses why the problem studied here may be of interest in other economic contexts, particularly in self referential macroeconomic learning models. Some concluding remarks are also presented here.

2. Optimality of memory length

Assume that a random variable μ_t evolves according to a first order auto-regressive process (AR(1)) as specified below

$(A.0)$ $\mu_{t+1} = \lambda\mu_t + (1-\lambda)\bar{\mu} + \varepsilon_t;\ 0 \leq \lambda < 1$

$(A.1)$ $\{\varepsilon_t\}$ is an i.i.d sequence with $E\varepsilon_t = 0; E\varepsilon_t^2 = \sigma_\varepsilon^2$.

$(A.2)$ μ_0 is given.

The unconditional (asymptotic) mean of the μ_t process is given by the constant $\bar{\mu}$. The true data generating process for μ_t is assumed unknown to the agents. On the other hand, agents need to forecast the current value of μ_t to make an economic decision. They forecast the time t realization of the random variable, μ_t, on the basis of the sample mean of the previous T data points, $\{\mu_{t-1}, \mu_{t-2,}\ .., \mu_{t-T}\}$. T is called the memory length of the agent. Call this forecast $\mu_t^e(T)$, which by definition is

$$\mu_t^e(T) = \frac{\sum_{i=1}^{T} \mu_{t-i}}{T} \tag{1}$$

Under rational expectations, agents would be assumed to know the true data generating process, an assumption which is usually considered implausible. Agents in this model deal with their lack of knowledge of the true structure by using a simple learning rule, which is essentially a variant of the least squares forecasting rule. However, even the simple learning rule considered here has much to be said in its favor. For one thing, (as

we shall presently show) this forecast is asymptotically unbiased for all memory lengths T. Secondly, the law of large numbers of Markov processes would imply that with large enough data the forecast would be expected to converge to the mean of the asymptotic distribution of the true process. But agents in this model are merely interested in forecasting the current realization of μ_t based on past data. However, even from this point of view of prediction, the forecast $\mu_t^e(T)$ has several attractive properties in the sense that it encompasses the optimal prediction for the important borderline cases of an i.i.d sequence (when $\lambda = 0$) and a random walk world (when $\lambda = 1$). If the true sequence is an i.i.d process, then $T \to \infty$ is optimal for prediction whereas if the true world is a random walk, then $T = 1$ is optimal for prediction (the best prediction in this case is given by the last period's value). However, in this section, we want to explore whether the choice of optimal T is affected when the true process is intermediate between these two extreme versions of the world (that is, when λ is between 0 and 1).

The forecast error made by the agent at any date t is given by $\mu_t^e(T) - \mu_t$. We first show that this forecast is unbiased for all memory lengths when the process has been in operation for a long period of time (i.e. asymptotically as $t \to \infty$).

Proposition 1. *The forecast error, $\mu_t^e(T) - \mu_t$, is asymptotically (i.e. as $t \to \infty$) unbiased for all T, that is, $\text{limit}_{t\to\infty} E(\mu_t^e(T) - \mu_t) = 0$.*

Proof. $\text{limit}_{t\to\infty} E(\mu_t^e(T) - \mu_t) = \text{limit}_{t\to\infty}(T^{-1}\sum_{i=1}^{T} E\mu_{t-i} - E\mu_t) = \bar{\mu} - \bar{\mu} = 0.$ ■

We now turn towards a characterizaton of the second moment properties of this forecast. A natural optimality criterion seems to be minimization of the mean squared error (MSE) of $\mu_t^e(T), E[(\mu_t^e(T) - \mu_t)^2]$. We assume that the process has been running for a long period of time so that $t \to \infty$ gives a reasonable approximation of this process. A natural advantage of this approximation is that it gets rid of the dependence of the optimal memory length on the initial condition of the process. This assumption is also in line with much of what is done in econometrics: one is usually interested in the statistical properties of estimators or predictors in the long run, that is, once the influence of the

initial conditions has died down. With this in mind, we assume that the agents want to minimize the *asymptotic* MSE, that is, want to minimize $\text{limit}_{t\to\infty}E[(\mu_t^e(T)-\mu_t)^2]$.[1] The basic choice problem of the agent is, therefore, to compute the memory length, T, which minimizes $\text{limit}_{t\to\infty}E[(\mu_t^e(T)-\mu_t)^2]$.

To economize on notation, let $E[(\mu_t^e(T)-\mu_t)^2]$ be denoted by $MSE_t^{est}(T)$ and $\text{limit}_{t\to\infty}$ $MSE_t^{est}(T)$ be denoted by $MSE_\infty^{est}(T)$. As a preliminary step we prove the following proposition.

Proposition 2. *For any $\lambda \in [0,1)$, we have*

$$MSE_\infty^{est}(T) = \sigma_\varepsilon^2[\frac{(1-\lambda)^2T(T+1) \ + \ 2(1-\lambda)\lambda^{T+1}T \ - \ 2\lambda(1-\lambda^T)}{(1-\lambda)^3(1+\lambda)T^2}]. \tag{2}$$

Proof. See Appendix A. ■

When $\lambda = 0$, $MSE_\infty^{est}(T)$ clearly decreases monotonically with T. However, when $\lambda > 0$, it is not immediately obvious from (2) as to how the expression behaves with T. To make this more transparent let us rewrite (2) in the following manner after rearranging terms

$$MSE_\infty^{est}(T) =$$

$$\sigma_\varepsilon^2\{\frac{1}{(1-\lambda^2)}(1+\frac{1}{T}) + \frac{2\lambda^{T+1}}{(1-\lambda)^2(1+\lambda)T} + \frac{2\lambda^{T+1}}{(1-\lambda)^3(1+\lambda)T^2} - \frac{2\lambda}{(1-\lambda)^3(1+\lambda)T^2}\}$$

This makes clear that while the first three terms within the curly brackets are indeed decreasing monotonically with T, the fourth term is increasing with T. Consequently, it is a question of which effect dominates. As a first step towards this analysis, I prove the following proposition.

Proposition 3. *For any $\lambda \in (0,1)$, $MSE_\infty^{est}(T)$ decreases monotonically with T for all $T \geq T(\lambda) = \frac{4\lambda}{(1-\lambda)^2}$.*

Proof. See Appendix B. ■

[1] Since the mean prediction error is asymptotically zero, this is also the asymptotic variance of prediction error.

To get an idea of the magnitude of $T(\lambda)$ for different λ, note that $T(.25) \approx 2$, $T(.5) = 8$, $T(.9) = 360$ and $T(.99) = 39,600$. It is easy to check that $T(\lambda)$ increases monotonically with λ. It follows from Proposition 3 that $\text{MSE}_\infty^{est}(T)$ decreases monotonically with T for all $T \geq 1$ provided λ is *small* enough. On the other hand, if λ is *large,* then Proposition 3 only guarantees that $\text{MSE}_\infty^{est}(T)$ decreases monotonically with T only for T *large enough* (specifically for $T \geq T(\lambda)$).

We sharpen Proposition 3 below.

Proposition 4. *For all* $\lambda \in [0, 0.5]$, $T \to \infty$ *minimizes* $MSE_\infty^{est}(T)$.

Proof. See Appendix C. ■

It will presently be shown that Proposition 4 is *not true* for all $\lambda \in [0, 1)$. In fact, in the proof of Proposition 4, it was shown that, for $\lambda > .5$, $T \to \infty$ can no longer be optimal since $T = 2$ has a smaller MSE. But can we actually compute the optimal memory length in this case? In fact we can prove the following:

Proposition 5. *For all* $\lambda \in (.5, .88]$, $T = 1$ *minimizes* $MSE_\infty^{est}(T)$.

Proof. See Appendix D. ■

The proof of Proposition 5 may lead one to suspect that $T = 1$ is optimal for all $\lambda \in (.5, 1)$. One can, in principle, look at values of λ arbitrarily close to 1 and solve the corresponding polynomial inequalities. However, the computation time increases very rapidly.[2] Instead I resorted to numerical simulations for values of λ close to 1 and found that the MSE indeed increases with T from $T = 1$ to $T = T(\lambda)$. Of course, Proposition 3 proves that the MSE must decrease for all $T \geq T(\lambda)$. So one can conjecture the following:

Conjecture: The optimal T is 1 for all $\lambda \in (.5, 1)$.

The broad picture that emerges then is that the optimal memory length is 1 when $\lambda \geq .5$ whereas it is infinity for $\lambda < .5$. One can, however, understand to some extent the intuition of these results. When $\lambda = 0$, μ_t is simply a sequence of i.i.d random variables

[2] To get an idea, it took a Pentium 233 Mhz *PC* with 192 MB of SDRAM almost three days to solve the polynomial inequalities up to $T = 250$ using *Mathematica Version 3.*

with mean $\bar{\mu}$. The MSE of prediction with memory T in this (correctly specified model) is given by $\sigma_\varepsilon^2(1+T^{-1})$ which is decreasing in T so that $T \to \infty$ minimizes this. So, for small λ, it may be reasonable to expect that large T will be optimal. At the other extreme, when $\lambda = 1$, μ_t follows a random walk so that the best prediction is given by the last realization, that is, $T = 1$ is optimal. So, for λ close to 1, it may be reasonable to expect that small T will be optimal. The striking thing that Propositions 4 and 5 tell us is that actually much more is true, namely, $T \to \infty$ is optimal for *all* $\lambda \in [0, 0.5]$ and $T = 1$ is optimal for *all* $\lambda \in (0.5, 0.88]$ (and perhaps even for all $\lambda \in (0.5, 1)$).

A related way to give some intuition for these results is the following. The autoregressive parameter of the AR(1) process (λ) may be interpreted to index the degree of mis-specification in the model given the agents' beliefs about the data generating mechanism. Suppose, for example, that agents believe they live in an i.i.d world and consequently use $T \to \infty$. If λ is close to 0, then the model is not too mis-specified and $T \to \infty$ continues to be optimal for prediction. On the other hand, when λ is close to 1, the model mis-specification is very severe and the use of more data for prediction is detrimental. Similarly, suppose that agents believe they live in a random walk world so that the best prediction of today's realization is simply the last period's value which is equivalent to using $T = 1$. The results of this section then show that $T = 1$ continues to be optimal when λ is close to 1 since the model is not too mis-specified. However, $T = 1$ is no longer optimal when λ is close to 0 since the model is heavily mis-specified.

3. ECONOMIC EXAMPLE MODELS

I now describe a couple of economic models which fit the framework of the problem studied in section 2.[3]

3.1. Profit maximization by the firm. In a way this example follows Muth (1961). Consider the problem of a firm choosing output in periods $t = 1, 2, 3...$ based on its forecast of the market prices for the respective periods. The realized price in period t is denoted by p_t. We assume that the price p_t follows an exogenous stochastic process. This would be appropriate in an open economy or for a monopolist facing infinitely elastic demand or

[3] The two example models are in fact borrowed in its entirety from Evans and Ramey (1998).

for a firm producing in a competitive market. In particular, assume that the price follows the process $p_t = \mu_t$ given by $(A.0)$, $(A.1)$, and $(A.2)$ with the added restriction that ε_t has a bounded support to ensure the non-negativity of price. The firm chooses output q_t at the end of period $t-1$ to maximize expected period t profits. Assuming quadratic costs $cq_t^2/2$, profits are given by

$$\Pi_t = p_t q_t - cq_t^2/2$$

so that expected profits are maximized by choosing $q_t = c^{-1}p_t^e$ where p_t^e is the expectation of p_t held by the firm at the end of period $t-1$. Instead of assuming rational expectations as Muth (1961) did, we assume that the forecast p_t^e is given by the sample mean of the previous T prices, i.e. by

$$p_t^e(T) = \frac{\sum_{i=1}^{T} p_{t-i}}{T}. \tag{3}$$

By using the optimal choice $q_t = c^{-1}p_t^e(T)$ profits may be rewritten as

$$\Pi_t(T) = (2c)^{-1}(2p_t p_t^e(T) - p_t^e(T)^2)$$

The firm wants to choose the T which maximizes $E\Pi_t(T)$.

On the other hand, consider the MSE of prediction of $p_t^e(T)$ which is given by $E[p_t^e(T) - p_t]^2$. Suppose now the firm instead chooses T to minimize $E[p_t^e(T) - p_t]^2 = E[p_t^e(T)^2 - 2p_t p_t^e(T) + p_t^2]$. However, since p_t is exogenous, this is equivalent to choosing T to minimize $E[p_t^e(T)^2 - 2p_t p_t^e(T)]$. Consequently, choosing T to minimize the MSE is equivalent to choosing T to maximize $E\Pi_t(T)$. We assume that the price process has been running for a long period of time so that the firm can ignore the effect of the initial conditons. The problem of the firm is, therefore, to choose the memory length T which maximizes $E\Pi_t(T)$ in the long run, that is, as $t \to \infty$. So the results of the previous section are applicable directly here.

Before proceeding any further I want to clarify some points which may be troubling the reader at this point. The first question relates to the utility of the results on optimal memory length in section 2. Given that the true price process is unknown to the firm, in what way are the results on optimal memory length useful to it? My defence here would

be the following. It may be fair to say that even if the true price process is not known *exactly* to the firm, it is quite likely to have *some* idea about the form of this process (say) after the conduct of some suitable econometric tests. For instance, these tests may lead the firm to entertain the possibility that the true price process is either a random walk or a process which is close to a random walk. This seems to me to be a particularly realistic situation given the notorious difficulties of econometric tests in distinguishing between a random walk world and a near random walk world (see Hamilton (1994) and also the discussion below on this point). The firm may know that $T = 1$ is optimal if the true world is a random walk. At the same time, it may conduct the optimality exercise of section 2 and conclude that even if the true world is only close to a random walk (say, the true value of λ is .9), it is still optimal to use $T = 1$ in its prediction. Given that the firm is uncertain about the true price process, the use of $T = 1$ in prediction can be defended on this ground (alone).

A related point can also be made here. The argument in the previous paragraph presupposes that the firm uses the forecast (3) in its prediction. An issue here may be the choice of the predictor (estimator) used by the firm. In section 2, I had presented several arguments as to why *a priori,* the firm may find the forecast (3) desirable to use on several grounds. These reasons ranged from (3) being an unbiased estimator of the true mean for all memory lengths to being expected to converge to the (true) mean for a large enough memory length for all values of λ. Even from the point of view of prediction, which is after all the main focus of the paper, this forecast encompasses the optimal prediction for the important borderline cases of an i.i.d sequence and a random walk world. However, more pertinently, a further case can be made here in its favor. In general, a predictor may function very well if the model is correctly specified whereas it may perform poorly if it is incorrectly specified. Arguably, the firm is unlikely ever to feel fully confident that it has the correct description of the real world. In these situations, the firm may prefer a *simple* predictor which performs (reasonably) well in a variety of circumstances rather than a predictor which performs extremely well in a correctly specified model but performs rather poorly in a mis-specified model. This provides an additional reason for the firm to

prefer the forecast (3). Perhaps a concrete example here will help to fix ideas. If the firm has rational expectations (RE), i.e. it knows the true form of the price process as well as the (correct) values of λ, $\bar{\mu}$ (and the variance of the unknown error term), then it has a MSE of σ_ε^2. Assume, without any loss of generality, that the true value of $\bar{\mu}$ is 0. The use of $T = 1$ in the forecast (3) yields an (asymptotic, i.e., as $t \to \infty$) MSE of $2(1+\lambda)^{-1}\sigma_\varepsilon^2$ which equals the MSE under RE if the true world is a random walk; otherwise it yields a higher MSE. For the purposes of comparison, we now consider another (plausible) predictor. Assume that the firm knows the true mean $\bar{\mu}$ and that price follows an AR(1) process. It, therefore, uses the predictor

$$p_t^{ar}(\lambda) = \lambda\mu_{t-1} \tag{4}$$

which depends on λ, assumed unknown to the firm. If the firm knew the true value of λ (say $\bar{\lambda}$) also (i.e. has RE), then it will attain higher expected profits than a firm using $T = 1$ in the forecast (3). However, if the firm incorrectly infers some value $\lambda \neq \bar{\lambda}$ on the basis of some statistical tests, then it may just as easily earn *smaller* expected profits with the use of the predictor (4) than with the use of (3) for a very wide range of values of λ. It is easy to check that if the firm infers (guesses) some value λ for the AR(1) parameter (possibly different from $\bar{\lambda}$), then the corresponding asymptotic (as $t \to \infty$) MSE associated with the predictor $p_t^{ar}(\lambda)$, (4), is

$$MSE(p_t^{ar}(\lambda)) = [1 + \frac{(\lambda - \bar{\lambda})^2}{1 - \bar{\lambda}^2}]\sigma_\varepsilon^2.$$

Obviously, $MSE(p_t^{ar}(\lambda)) = \sigma_\varepsilon^2$ if $\lambda = \bar{\lambda}$. But $MSE(p_t^{ar}(\lambda))$ will be more than the MSE for the forecast (3) with $T = 1$, $MSE_\infty^{est}(1)$, for a wide range of values of λ. For example, one can check that $MSE(p_t^{ar}(\lambda)) > MSE_\infty^{est}(1)$ for all $\lambda < .98$ if $\bar{\lambda} = .99$; as well as for all $\lambda < .9$ if $\bar{\lambda} = .95$, and for all $\lambda < .8$ if $\bar{\lambda} = .9$. Thus, if $\bar{\lambda} = .99$, then the predictor (4) fares worse than (3) for all $\lambda < .98$ and only performs better otherwise. The firm can, therefore, earn higher expected profits with the use of the simple predictor (3) than with the use of the predictor (4) if it incorrectly infers the value of the AR(1) parameter (even though it knows the true mean $\bar{\mu}$) for a wide range of λ. The obvious question which arises now is how likely is it that the firm may incorrectly infer the value of the AR(1) parameter

λ on the basis of statistical tests? The answer is that this is very likely for values of λ close to or equal to 1. There is an extensive literature in econometrics that discusses the difficulties in making a correct inference in this situation.[4] For example, Evans and Savin (1981) provide the power functions for a test of the (null) hypothesis of $\lambda = 1$ for various sample sizes for the AR(1) process considered here with $\bar{\mu} = 0$, which is assumed known to the investigator (firm). At $\lambda = .9$, a sample of size 100 only achieves a power of 56% whereas at $\lambda = .99$, a sample as large as 400 merely achieves a power of 12.8%.[5] The situation is similar for a test of the hypothesis of stationarity. For instance, Evans and Savin (1981) find that the power functions for testing the hypothesis of $\lambda = .95$ continues to be poor. With a sample of size 100, the power is 13% at $\lambda = .9$ and only 60% at $\lambda = 1$. Given this situation, the firm may be quite content to use the forecast (3) since, in a sense, this protects it from a range of model mis-specification which the more (sophisticated!) predictor (4) is unable to. Furthermore, in conclusion, one can add that the forecast (3) has the advantage that the optimal memory length is invariant to a range of values of λ- for example, the optimal memory length is 1 for all $\lambda > .5$. Consequently, the firm need not worry too much about the inadequacy of econometric tests in distinguishing between random walk and near random walk processes if it uses the forecast (3) in its prediction.[6]

3.2. Permanent Income Hypothesis. This corresponds to the first example in Lucas (1976). Consumption is given by

$$c_t = c_{pt} + u_t$$

$$c_{pt} = k y_{pt}$$

$$y_{pt} = (1-\delta)\sum_{i=0}^{\infty} \delta^i y^e_{t+i},\ 0 < \delta < 1.$$

Here u_t is a white noise process denoting transitory consumption. c_{pt} denotes permanent consumption, y_{pt} denotes permanent income, δ is the household's discount factor and y^e_{t+i}

[4] For a sample of this literature, see Evans and Savin (1981, 1984) and Dickey and Fuller (1979, 1981).

[5] See their Table 4, p. 771, for the details. The situation is (obviously) worse if the firm does not even know the true mean $\bar{\mu}$. For a sample of the power functions of the random walk hypothesis in the latter case, see Table 6, p. 1260, of Evans and Savin (1984).

[6] Needless to say, most of these arguments are also valid for a firm uncertain about the price process for values of λ equal to or close to zero.

is the household's time t forecast of income at time $t+i$, y_{t+i}. The income process y_t is now assumed to follow an AR(1) process ($y_t = \mu_t$ here). While Lucas (1976) assumed rational expectations, we instead assume that $y^e_{t+i} = y^e_t(T)$ for all $i = 0, ..\infty$ so that y_{pt} $= y^e_t(T)$ and $c_t = k y^e_t(T) + u_t$. $y^e_t(T)$ is given the sample mean of the last T realizations, that is,

$$y^e_t(T) = \frac{\sum_{i=1}^{T} y_{t-i}}{T}$$

The agent's problem now is to choose the memory T which minimizes the MSE of prediction, $E[y^e_t(T) - y_t]^2$. We again assume that the income process y_t has been evolving for a long period of time so that the appropriate problem of the agent is to choose the memory length which minimizes $E[y^e_t(T) - y_t]^2$ in the long run, that is, as $t \to \infty$. Our results in section 2 then suggest that if $\lambda < 0.5$, the agent should use as much data as possible to minimize the MSE whereas he should use $T = 1$ if $\lambda > 0.5$. Moreover, an accurate forecast of $y^e_t(T)$ provides him with an accurate forecast of permanent income (y_{pt}) and, therefore, of (permanent) consumption (c_{pt} or c_t). Note that agents are using the same forecast, $y^e_t(T)$, for making predictions over longer horizons. This is fully rational if the true process, y_t, is given by a constant (unknown to the agent) plus some i.i.d noise. In this case, the $l-$ step ahead forecast (y^e_{t+l}) from time origin t would be given by $y^e_t(T)$ (see Abraham and Ledolter (1983)). This forecast is unbiased and has mean squared prediction error $\sigma^2_\varepsilon(1+\frac{1}{T})$ which is obviously decreasing in T. Consequently, the use of this forecast on the part of agents can be rationalized by assuming that they underparametrize the true Markovian process to be an i.i.d sequence. Another way to rationalize the choice of the same forecast, $y^e_t(T)$, for making predictions over longer horizons is to assume that agents believe they live in a random walk world and, therefore, use $T = 1$. In this case, the optimal $l-$ step ahead linear forecast (y^e_{t+l}) would be given by $y^e_t(1)$ (see Hamilton (1994)).

4. CONSISTENCY OF EXPECTATIONS

In this section, I study, through an illustrative example, whether an examination of forecast errors made by agents when they use the forecast (1) would cause them to suspect a mis-specification in their model. When agents are using a mis-specified model, it may

not be possible for them to learn the rational expectations equilibrium (REE). Then the question is whether the persistent prediction errors they make (errors which do not vanish asymptotically) show any kind of systematic pattern. For concreteness, I take the example of the producer forecasting prices. I assume that the price process is given by $p_{t+1} = \mu_t p_t(1-p_t)$ where μ_t is AR(1) with the noise having bounded support.[7] The producer, on the other hand, assumes that the price process is $p_{t+1} = (\mu + v_{t+1})p_t(1-p_t)$ with $\{v_t\}$ being an i.i.d sequence and forecasts this price by $\mu_t^e p_t(1-p_t)$ where μ_t^e is given by (1). The producer uses data on prices to test for mis-specification.

There are various ways to test for mis-specifications. The most obvious one (often suggested by econometricians as a first step) is a plot of the residuals over time. Another step which Harvey (1989) advocates is a plot of the residuals against one of the explanatory variables (here price) which may be done if agents suspect some functional mis-specification. Simulations indicate that such a plot of the residuals in this model does not reveal any mis-specification. This should not seem that surprising since after all the producer has the correct functional form of the evolution of prices. On the other hand, Bray and Savin (1986) in their analysis of learning in a cobweb model checked whether agents would detect any mis-specification in their model through some diagnostic checks. For instance, they conducted tests for parameter constancy. For this example, however, since the producer does assume a moving parameter there is no reason for him to conduct tests to check for parameter constancy.

I will instead take a different route which has recently been advocated by Hommes (1998) and Hommes and Sorger (1998). They propose the concept of *consistent expectations equilibria* (CEE) which requires that agents correctly perceive all autocorrelations of the process. As Evans and Honkapohja (1999) note, this makes it a very stringent criteria. I will now examine whether the expectations of the producers are consistent or not. Hommes (1998) defines consistency of expectations in terms of the autocorrelation function (ACF) of the expectational errors. In our case these will be the prediction errors

[7] Note that the price process being given by the logistic map could potentially be complicated. However, in my simulations, I let the initial condition of μ_t be in the region where the steady state is the global attractor. Consequently, the failure of agents to detect any mis-specification is not due to any *chaotic* pattern in the prediction errors.

denoted by $e_t(T)$. The (empirical) ACF ρ_k of $e_t(T)$ is defined as

$$\rho_k = \frac{c_k}{c_0} \; ; \; -1 \leq \rho_k \leq 1$$

$$c_k = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N-k} (e_t(T) - e^-(T))(e_{t+k}(T) - e^-(T))$$

$$\text{with } e^-(T) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} e_t(T)$$

Hommes (1998) defines expectations to be consistent if the autocorrelation coefficients ρ_k of the expectational errors are zero for all $k \geq 1$. He defines expectations to be *weakly* consistent if there exists a $K \geq 2$ such that the autocorrelation coefficients ρ_k of the expectational errors are zero for all $k \geq K$. Expectations are defined to be inconsistent if they are not weakly consistent. Agents having inconsistent expectations would have ample cause to believe that their model is mis-specified.

Now let us turn to the question of testing this definition on our producers. I first consider small values of λ. Since for small values of λ (precisely for $\lambda \leq .5$) the MSE is decreasing in T, it is optimal to use as much data as possible for predicton. So let us examine whether the autocorrelations in prediction errors are significantly different from zero for large T in this case. Simulations indicate that the ACFs are insignificantly different from zero for values of λ close to zero; so expectations are consistent. Figure 1a plots the (normalized) least squares prediction residuals for 80 observations after dropping the first 300 transients when $\lambda = .1, \bar{\mu} = 2$, $T = 50$ and the noise is uniform with support $[-10^{-3}, 10^{-3}]$. The residuals do not seem to have any systematic pattern. Figure 1b plots the corresponding sample ACFs at the first 20 lags in the above case. The straight lines have a height of $\pm\frac{2}{\sqrt{M}}$ where M is the sample size. Only ACFs above or below the straight lines would be considered insignificantly different from zero at the 5% level. As is clear from the figure, all ACFs are insignificantly different from zero so that expectations are consistent. A similar picture emerges for a lot of values of λ less than 0.5. The results also do not seem to be sensitive to the distribution of the noise or to its magnitude.

For $\lambda > .5$, we have seen that $T = 1$ is optimal. The relevant question now is whether agents using the optimal memory will suspect some mis-specification in their model. As an

illustration, Figure 2a plots the (normalized) least squares residuals for 35 observations after dropping the first 300 transients when $\lambda = .6, \bar{\mu} = 2$, $T = 1$ and the noise is uniform with support $[-10^{-3}, 10^{-3}]$. The residuals look quite random. Figure 2b plots the corresponding sample ACFs at the first 20 lags in the above case. Expectations are again consistent.

As another example, Figure 3a plots the (normalized) least squares residuals for 35 observations when $\lambda = .8, \bar{\mu} = 2$, $T = 1$ and the noise is uniform with support $[-10^{-3}, 10^{-3}]$. The residuals are again seemingly random. Figure 3b plots the corresponding sample ACFs at the first 20 lags in the above case. None of the autocorrelation coefficiants differ from zero significantly. More generally, the same type of picture emerges if agents are using a T which is close to (but not necessarily) 1. However, the picture changes if agents use a large T in this case. For example, figures 4a and 4b plots the (normalized) residuals and the sample ACFs when $\lambda = .6$ and $T = 50$ for 100 observations. Note that the ACFs at the first 2 lags are significant with the one at the first lag strongly so.

Collecting these observations together, the broad theme that emerges is the following. If agents use the optimal T, expectations are in most cases consistent. If they use a T which is reasonably close to the optimal T, then expectations are at least weakly consistent. If they use a T which is far from the optimal T, then even if expectations are weakly consistent, the ACF at the first lag is often rejected strongly so that they may suspect a mis-specification in their model. This may in turn lead them to conduct more sophisticated econometric tests to detect the source of the mis-specification.

5. Discussion and Concluding Remarks

The paper has considered scenarios where the exogenous variable follows a stochastic process unknown to the agent. If the true data generating process is unknown (and potentially complex), economic agents may be expected to use simple underparametrized representations of the process to make their forecasts. They can then obtain the best forecast within this class. An appropriate bounded rationality assumption seems to be that agents, in the terminology of Sargent (1999, Ch. 6), have "optimal misspecified beliefs". A similar idea has been explored in this paper where agents forecast the current

value by using a version of the least squares forecast with a fixed amount of data. This forecast embodies the optimal forecast for two extreme versions of the world (that is, when the true world is i.i.d and when it is a random walk). Agents would prefer a forecasting rule which is robust to a mis-specification in the model. It has been found that such a robust choice of the memory length does exist for the class of models explored here.

The basic idea explored in the paper is related to some recent studies in the macroeconomic learning literature in a stochastic setup. Broadly speaking, the question analyzed here falls within the spectrum of learning in mis-specified models. Recent studies which explore similar ideas include Sargent (1999), Evans and Honkapohja (1999, 2000), and Hommes and Sorger (1998). As an illustration, let us consider Evans and Honkapohja (2000). The idea they discuss is the following. In the literature the specification of the agents' learning rule comes (usually) from the underlying rational expectations equilibrium (REE) of the economy. Thus, for example, if the macroeconomic model has a REE which takes the form of an i.i.d sequence, then the agents' learning rule (the perceived law of motion or PLM for short) would also be an i.i.d sequence with unknown parameter. Thus, in a certain sense, the PLM of agents who are learning and the actual law of motion (ALM) sit in the same functional space so that with right parameter values the PLM coincides with the REE of interest. Consequently, even though the model of the agents is mis-specified while they are learning, there is a possibility of learning being *complete* in the sense that the economy settles to an REE if the learning dynamics converges. In general, however, there is no reason why this should be the case. As emphasized by Evans and Honkapohja (2000), economic agents, like econometricians, may fail to correctly specify the ALM even asymptotically. They show that such mis-specifications can radically alter both the nature of the equilibria as well as the stability conditions for convergence to such equilibria. For example, they examine a version of the Cagan inflation model with lagged endogenous variables which has two minimal state variable solutions of the AR(1) form under rational expectations. If the agents' PLM takes the form of an i.i.d process, then the equilibrium under such dynamics (termed *restricted perceptions equilibrium* by

them[8]) becomes unique. In general, therefore, the form of the PLM could crucially affect the nature and stability of equilibria.[9]

Hommes and Sorger (1998), on the other hand, consider models where the PLM is linear but the ALM is nonlinear. They introduce the concept of *consistent expectations equilibrium* (CEE) by the property that the PLM and ALM are indistinguishable in terms of the sample average and sample autocorrelations of the observed variable. Their emphasis is on the fact that agents should not be able to detect any mis-specification in their model on the basis of simple statistical tests. This point is particularly relevant for macroeconomic learning models since the model agents use is mis-specified during the transition to REE even though asymptotically it may be correctly specified.

The prevailing literature has not emphasized the size of memory to be an important issue for learning models. However, the current paper shows that this could potentially be important. In reality, the form of the PLM and the size of the memory could be intimately related which in turn can affect the nature and stability of equilibria. I now want to add some of my thoughts on this subject by means of a simple illustrative example. Consider the class of models given by

$$y_t = \alpha + \beta E^*_{t-1} y_t + v_t$$

with $\beta \neq 1$, which could describe, for example, the Lucas (1973) "island" model combining a "surprise" aggregate supply function and a "quantity theory" demand equation with y_t being interpreted as the price and v_t being an i.i.d noise.[10] We use the same notation $E^*_{t-1} y_t$ to denote the expectations of agents under both rational expectations and learning. This model has a unique REE given by

$$y_t = \bar{a} + v_t$$

where $\bar{a} = \alpha/(1-\beta)$. Under learning, agents are assumed not to know $\bar{a}$ but have a PLM which corresponds to the REE. Therefore, they estimate $\bar{a}$ by the sample mean of the

[8] The restricted perceptions equilibrium concept is also closely related to the notion of *reduced order limited information* rational expectations equilibria (REE) introduced in Sargent (1991).

[9] An application of this idea is in Sargent (1999) where he suggests that a similar form of incomplete learning may be an essential ingredient in the rise and decline of inflation in post-war America.

[10] See Evans and Honkapohja (1999) for a more detailed explanation as to how this reduced form arises.

data (given by $a_t = t^{-1}\sum_{i=0}^{t-1} y_i$) which also happens to be their forecast, $E^*_{t-1}y_t$, in this case. This means that the ALM followed by y_t will actually be

$$y_t = \alpha + \beta a_t + v_t.$$

Asymptotically, with large enough data, the forecasts of agents can be shown to converge to the REE if $\beta < 1$. On the other hand, the agents' forecasting model is mis-specified during the transition to REE and (as should be evident from the ALM) this mis-specification could potentially be quite severe. Consequently, it is possible that agents might abandon their PLM during the transition process in which case convergence to the REE may not take place. However, I think that the analysis of the paper provides an added justification as to why agents might stick with their learning rule in such a scenario. The learning rule with infinite data can be optimal not only when the process is i.i.d but also when the process is auto-correlated. The robustness of this rule to mis-specifications in the agents' model make it more likely for agents to stick with it.[11]

Now suppose, for whatever reason, the PLM of agents is a random walk (without drift). In the spirit of the learning literature, agents assume the actual process followed by the economy to be time invariant and choose a learning procedure which is consistent with their PLM. In this case, the optimal forecast would be to use the last period's value, that is, $E^*_{t-1}y_t = y_{t-1}$ (this is equivalent to using $T = 1$). Now the ALM becomes

$$y_t = \alpha + \beta y_{t-1} + v_t \tag{5}$$

which is an AR(1) process. To draw a parallel with the process discussed in the paper, let us assume, without any loss of generality, that $\alpha = (1-\lambda)\bar{\mu}$ and $\beta = \lambda$. λ can now be interpreted to index the influence of expectations in the model. If expectations matter a lot (that is, λ is close to 1), then the PLM and the ALM may not be distinguishable based on a finite number of observations.[12] Thus, this may be considered a more plausible description of the world where agents are learning since any mis-specification may be hard to detect. This would also be more in the spirit of the ideas discussed by Hommes and Sorger (1998)

[11] Of course, more work needs to be done to complete the argument since the ALM is not AR(1) in this case.

[12] See the discussion in Section 3.1 above and Hamilton (1994).

through their idea of consistent expectations. Again, even if agents are unsure about their exact model or suspect some slight mis-specification in their model, they are more likely to stick with the choice of $T = 1$ since this choice has been demostrated to be robust to a mis-specification in the agents' model.[13] Therefore, this provides an added justification as to why (5) will provide a more plausible description of the "learning" economy. Of course, convergence to the unique REE does not take place in this case since the ALM is AR(1). In this way, the size of the memory may affect the stability of REE.

The question of optimality of memory length which was the main thrust of this paper is, however, somewhat related to the ideas explored in Evans and Honkapohja (1993) and Sargent (1999). In Evans and Honkapohja (1993), and more generally in the statistical and engineering literature, when agents suspect some structural change (or time varying parameter) the advice given to them is to use a "constant gain" instead of "decreasing gain" in their learning algorithm. This essentially means that instead of putting decreasing weight to current observations (so that asymptotically the weight vanishes as in least squares estimation, for example), one puts some constant weight to current data. This procedure of constant gain involves a trade-off between bias and variance when used to adapt to an exogenous time-varying process. A larger value of the gain reduces the bias but increases the variance of the forecast. Evans and Honkapohja (1993) examine the question of the optimal gain parameter to use for an agent in the context of an overlapping generations economy. They furthermore examine whether there exists an equilibria in learning rules in the sense that no agent has an incentive to deviate from his choice of the gain (parameter) given the gain (parameter) of all other agents. A similar idea is explored in Sargent (1999 Ch. 6) where he takes Bray's (1982) model and instead of assuming that the forecast is given by the sample mean of the observations (as Bray did), he assumes that forecasts are formed adaptively with a fixed gain parameter C. The agent then chooses C to minimize the one-step ahead forecasting error. Such specifications can alter the nature of equilibria in interesting ways.

As I have tried to indicate, this paper opens up further avenues of research. One

[13] In fact, in this case, $T = 1$ will be optimal both with respect to the PLM and the ALM if $\lambda > 0.5$.

would expect finite memory learning rules to alter the nature of equilibria in stochastic self-referential models. As part of an ongoing project, I am studying this question in collaboration with Seppo Honkapohja (see Honkapohja and Mitra (1999)). One can also try to analyze questions of equilibria in learning rules (in terms of T) in the sense discussed in the previous paragraph for self-referential models.

6. References

Abraham B. and Ledolter J. (1983):" Statistical Methods for Forecasting".

Bray, M.M. (1982):" Learning, Estimation and the Stability of rational expectations" *Journal of Economic Theory*, 26 #2, 318-339.

Dickey, D.A. and Fuller , W.A. (1979):" Distribution of the Estimators for Autoregressive Time Series with a Unit Root", *Journal of the American Statistical Association*, 79 #366, p. 427-431.

Dickey, D.A. and Fuller , W.A. (1981):" Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root", *Econometrica*, 49 #4, p. 1057-1072.

Evans, G.W. and Honkapohja, S. (1993): "Adaptive Forecasts, Hysteris, and Endogenous Fluctuations". *Economic Review Federal Reserve Bank of San Francisco* Number 1.

Evans, G.W. and Honkapohja, S. (1999): " Learning Dynamics," in Taylor and Woodford, Handbook of Macroeconomics , Volume 1 chapter 7.

Evans, G., and S. Honkapohja. (2000). *Learning and Expectations in Macroeconomics.* Princeton, New Jersey: Princeton University Press, forthcoming.

Evans, G.W. and Ramey, G. (1998): "Adaptive Expectations, Underparametrization and the Lucas Crique", mimeo.

Evans, G.B.A. and Savin, N.E. (1981): "Testing for unit roots: 1", *Econometrica*, 49 #3, p. 753-779.

Evans, G.B.A. and Savin, N.E. (1984): "Testing for unit roots: 2", *Econometrica*, 52 #5, p. 1241-1269.

Hamilton, J. (1994): Time Series Analysis. Princeton University Press.

Harvey, A. C.(1989): Forecasting, structural time series models and the Kalman Filter.

Hommes, C. H. (1998): "On the conistency of backward-looking expectation The case of the cobweb," *Journal of Economic Behavior & Organization, 33, 333-362.*

Hommes, C. H. and Sorger G. (1998): "Consistent Expectations Equilibria," *Macroeconomic Dynamics 2, 287-321.*

Honkapohja S. and Mitra K. (1999): "Learning with bounded memory in stochastic models," University of Helsinki Discussion Paper No. 456.

Lucas, R. E. (1973): "Some international evidence on output-inflation tradeoffs" *American Economic Review*, 63, 326-334.

Lucas, R. E. (1976): "Econometric Policy Evaluation: A Critique", *Journal-of-Monetary-Economics*; 1(2), Supplementary Series 1976, pages 19-46.

Muth, J. (1961): "Rational Forecasts and the theory of price movements" *Econometrica*, 29, 315-335.

Sargent, T. J. (1991): "Equilibrium with signal extraction from endogeneous variables," *Journal of Economic Dynamics and Control, 15, 245-273.*

Sargent, T. J. (1999): The Conquest of American Inflation. Princeton University Press.

A. Proof of Proposition 2

First, recall that

$$\mu_t^e(T) = \frac{\sum_{i=1}^{T} \mu_{t-i}}{T}$$

Next observe that for all i, $1 \leq i \leq T-2$, we can write

$$\mu_{t-i} = \lambda^{T-i}\mu_{t-T} + \mu^-(1-\lambda)(1+\lambda+\lambda^2+...+\lambda^{T-i-1})+$$

$$\varepsilon_{t-i-1} + \lambda\varepsilon_{t-i-2} + ... + \lambda^{T-i-2}\varepsilon_{t-T+1} + \lambda^{T-i-1}\varepsilon_{t-T}$$

The previous line simply involves writing μ_{t-i} in terms of μ_{t-T} and the intermediate error terms. Using this fact we can show that

$$\sum_{i=1}^{T} \mu_{t-i} =$$

$$\mu_{t-T} + \{\lambda\mu_{t-T} + \mu^-(1-\lambda) + \varepsilon_{t-T}\}+$$

$$\{\lambda^2\mu_{t-T} + \mu^-(1-\lambda)(1+\lambda) + \varepsilon_{t-T+1} + \lambda\varepsilon_{t-T}\}+$$

$$\{\lambda^3\mu_{t-T} + \mu^-(1-\lambda)(1+\lambda+\lambda^2) + \varepsilon_{t-T+2} + \lambda\varepsilon_{t-T+1} + \lambda^2\varepsilon_{t-T}\} + ..+$$

$$\{\lambda^{T-1}\mu_{t-T} + \mu^-(1-\lambda)(1+\lambda+\lambda^2+..+\lambda^{T-2})+$$

$$\varepsilon_{t-2} + .. + \lambda^{T-3}\varepsilon_{t-T+1} + \lambda^{T-2}\varepsilon_{t-T}\}$$

$$= \mu_{t-T}(1+\lambda+\lambda^2+..+\lambda^{T-1})+$$

$$\mu^{-}(1-\lambda)\{1+(1+\lambda)+(1+\lambda+\lambda^2)+..+(1+\lambda+\lambda^2+..+\lambda^{T-2})\}+$$

$$\varepsilon_{t-T}\{\frac{(1-\lambda^{T-1})}{(1-\lambda)}\}+\varepsilon_{t-T+1}\{\frac{(1-\lambda^{T-2})}{(1-\lambda)}\}+..+\varepsilon_{t-3}\{\frac{(1-\lambda^{2})}{(1-\lambda)}\}+\{\frac{(1-\lambda)}{(1-\lambda)}\}\varepsilon_{t-2}$$

$$=\mu_{t-T}(\frac{1-\lambda^{T}}{1-\lambda})+\mu^{-}(1-\lambda)(\frac{(1-\lambda)T-1+\lambda^{T}}{(1-\lambda)^2})+\varepsilon_{t-T}\{\frac{(1-\lambda^{T-1})}{(1-\lambda)}\}+\varepsilon_{t-T+1}\{\frac{(1-\lambda^{T-2})}{(1-\lambda)}\}+$$

$$.....+\varepsilon_{t-3}\{\frac{(1-\lambda^{2})}{(1-\lambda)}\}+\varepsilon_{t-2}\{\frac{(1-\lambda)}{(1-\lambda)}\}$$

Moreover since

$$\mu_t=\lambda^{T}\mu_{t-T}+\mu^{-}(1-\lambda^{T})+\varepsilon_{t-1}+\lambda\varepsilon_{t-2}+....+\lambda^{T-1}\varepsilon_{t-T}$$

the error made in prediction is eventually given by

$$\mu_t^{e}(T)-\mu_t=(\mu_{t-T}-\mu^{-})\{\frac{1-\lambda^{T}-(1-\lambda)\lambda^{T}T}{(1-\lambda)T}\}+\varepsilon_{t-T}\{\frac{(1-\lambda^{T-1})}{(1-\lambda)T}-\lambda^{T-1}\}+$$

$$\varepsilon_{t-T+1}\{\frac{(1-\lambda^{T-2})}{(1-\lambda)T}-\lambda^{T-2}\}+....+\varepsilon_{t-2}\{\frac{(1-\lambda)}{(1-\lambda)T}-\lambda\}-\varepsilon_{t-1}$$

The MSE of the predictor ($MSE_t^{est}(T)$) is, therefore, given by

$$MSE_t^{est}(T):=E[(\mu_t^{e}(T)-\mu_t)^2]=\{\frac{1-\lambda^{T}-(1-\lambda)\lambda^{T}T}{(1-\lambda)T}\}^2E[(\mu_{t-T}-\mu^{-})^2]+$$

$$\sigma_{\varepsilon}^2[\{\frac{(1-\lambda^{T-1})}{(1-\lambda)T}-\lambda^{T-1}\}^2+\{\frac{(1-\lambda^{T-2})}{(1-\lambda)T}-\lambda^{T-2}\}^2+....+\{\frac{(1-\lambda)}{(1-\lambda)T}-\lambda\}^2+1)]$$

μ_t, being an AR(1) process, has a stationary distribution asymptotically (i.e. as $t\to\infty$) with variance $\frac{\sigma_{\varepsilon}^2}{(1-\lambda^2)}$, that is, $limit_{t\to\infty}E[(\mu_{t-T}-\mu^{-})^2]=\frac{\sigma_{\varepsilon}^2}{(1-\lambda^2)}$. Consequently, as $t\to\infty$, the expression for the MSE simplifies further to

$$MSE_{\infty}^{est}(T)=$$

$$\frac{\sigma_\varepsilon^2}{(1-\lambda^2)}\{\frac{1-\lambda^T-(1-\lambda)\lambda^T T}{(1-\lambda)T}\}^2 +$$

$$\sigma_\varepsilon^2[\{\frac{(1-\lambda^{T-1})}{(1-\lambda)T}-\lambda^{T-1}\}^2+\{\frac{(1-\lambda^{T-2})}{(1-\lambda)T}-\lambda^{T-2}\}^2+....+\{\frac{(1-\lambda)}{(1-\lambda)T}-\lambda\}^2+1]$$

$$=\sigma_\varepsilon^2[\frac{(1-\lambda)^2T(T+1)\ +\ 2(1-\lambda)\lambda^{T+1}T\ -\ 2\lambda(1-\lambda^T)}{(1-\lambda)^3(1+\lambda)T^2}].$$

In the final line above I have simply noted the end result which follows after simplification of the expression in the previous line and collecting all the terms involving σ_ε^2.

B. Proof of Proposition 3

Differentiating $MSE_\infty^{est}(T)$ with respect to T, we get

$$\frac{d[MSE_\infty^{est}(T)]}{dT}=\frac{A\sigma_\varepsilon^2}{(1-\lambda)^3(1+\lambda)T^3} \text{ where}$$

$$A:=-2\lambda(1-\lambda-\ln\lambda)\lambda^T T-(1-\lambda)^2T+2\lambda(1-\lambda)(\ln\lambda)\lambda^T T^2+4\lambda(1-\lambda^T)$$

The sign of A, which depends on both T and λ, determines whether MSE is increasing or decreasing with T. Observe that

$$A=-2(1-\lambda-\ln\lambda)\lambda^{T+1}T-(1-\lambda)^2T+2(1-\lambda)(\ln\lambda)\lambda^{T+1}T^2+4\lambda(1-\lambda^T)$$
$$<4\lambda(1-\lambda^T)-(1-\lambda)^2T$$

The strict inequality follows since the first and third terms are negative *for all* T and λ. So continuing

$$A<4\lambda(1-\lambda^T)-(1-\lambda)^2T<4\lambda-(1-\lambda)^2T$$

Define $T(\lambda)=\dfrac{4\lambda}{(1-\lambda)^2}$.

Then, from the above string of inequalities, it follows that $A<0$ for all $T>T(\lambda)$. When $T=T(\lambda)$, it is also easy to check that $A<0$. This proves the proposition.

C. Proof of Proposition 4

First let $\lambda^* = \frac{\sqrt{37}-1}{12} \approx .424$.[14] λ^* is the (unique) value of λ at which $MSE_\infty^{est}(2) = MSE_\infty^{est}(3)$. Then $T(\lambda^*) \approx 5.1$. By Proposition 3, we know that $\text{MSE}_\infty^{est}(T)$ decreases with T for all $T \geq 6$. The question is what happens for $T \leq 5$. To answer this, first observe the following:

$MSE_\infty^{est}(\ T+1) > MSE_\infty^{est}(\ T)$ if and only if (iff)

$$\frac{(1-\lambda)^2(T+1)(T+2)\ +\ 2(1-\lambda)\lambda^{T+2}(T+1)\ -\ 2\lambda(1-\lambda^{T+1})}{(T+1)^2} >$$

$$\frac{(1-\lambda)^2T(T+1)\ +\ 2(1-\lambda)\lambda^{T+1}T\ -\ 2\lambda(1-\lambda^{T})}{T^2}$$

iff

$$(1-\lambda)^2(\frac{1}{T+1}-\frac{1}{T})+2(1-\lambda)(\frac{\lambda^{T+1}}{T})(\frac{\lambda T}{T+1}-1)-2\lambda(\frac{1}{(T+1)^2}-\frac{1}{T^2})+\frac{2\lambda^{T+1}}{T^2}(\frac{\lambda T^2}{(T+1)^2}-1) > 0$$

The last line shows that, for given T, the above inequality is a polynomial in λ. One can verify that for all $\lambda \leq \lambda^*$,[15]

$$MSE_\infty^{est}(6) < MSE_\infty^{est}(5) < MSE_\infty^{est}(4) < MSE_\infty^{est}(3) \leq MSE_\infty^{est}(2) < MSE_\infty^{est}(1).$$

This proves that the $\text{MSE}_\infty^{est}(T)$ decreases with T for all $\lambda \leq \lambda^*$.

Before proceeding, note that

$$MSE_\infty^{est}(1) < MSE_\infty^{est}(2) \text{ iff } \frac{2}{(1+\lambda)} < \frac{3+2\lambda}{2(1+\lambda)} \text{ which is true iff } \lambda > .5.$$

Now consider the case when $\lambda \in (.424, .428]$.[16] Proposition 3 tells us that the MSE decreases with T for all $T > 5$ in this interval of λ. It is also possible to verify that for all $\lambda \in (.424, .428]$, $MSE_\infty^{est}(6) < MSE_\infty^{est}(5) < MSE_\infty^{est}(4) \leq MSE_\infty^{est}(3)$ and $MSE_\infty^{est}(2) <$

[14] Henceforth, all values of λ will be rounded off to the third decimal place.

[15] I used the "Inequality Solve" package in *Mathematica* Version 3.0 to solve algebraically for these and all of the succeeding polynomial inequalities which appear in the proofs.

[16] The right hand number of this interval, .428, is the (unique) value of λ (rounded off to the third decimal place) at which $MSE_\infty^{est}(3) = MSE_\infty^{est}(4)$.

$MSE_{\infty}^{est}(3)$. Since we already know that $MSE_{\infty}^{est}(2) < MSE_{\infty}^{est}(1)$, the optimal T can be computed by comparing $MSE_{\infty}^{est}(2)$ and $MSE_{\infty}^{est}(T \to \infty)$. On comparing we get

$$MSE_{\infty}^{est}(T \to \infty) < MSE_{\infty}^{est}(2)$$

$$\text{iff } \frac{1}{(1-\lambda)(1+\lambda)} < \frac{3+2\lambda}{2(1+\lambda)}$$

$$\text{iff } 2\lambda^2 + \lambda - 1 < 0$$

iff $\lambda < .5$ (the negative root being inadmissible). This proves that the optimal $T \to \infty$ when $\lambda \in (.424, .428]$.

At the risk of being repetitious, consider now the interval of $\lambda \in (.428, .446]$.[17] In this case, Proposition 3 tells us that the MSE decreases with T for all $T \geq 6$. The solution of the successive polynomial inequalities show that $MSE_{\infty}^{est}(6) < MSE_{\infty}^{est}(5) \leq MSE_{\infty}^{est}(4)$ as well as that $MSE_{\infty}^{est}(2) < MSE_{\infty}^{est}(3) < MSE_{\infty}^{est}(4)$. We already know that $MSE_{\infty}^{est}(2) < MSE_{\infty}^{est}(1)$. This means that the optimal T can again be computed by comparing $MSE_{\infty}^{est}(2)$ with $MSE_{\infty}^{est}(T \to \infty)$ which has been done above so that the optimal $T \to \infty$.

In a similar fashion consider neighbouring intervals like $(.446, .466]$, $(.466, .486]$, and $(.486, .5]$. Note that all these intervals arise out of comparing the MSE associated with adjacent memory lengths. For all of these intervals, the optimal T can, as before, be shown to be $T \to \infty$. This proves the proposition.

D. Proof of Proposition 5

When $\lambda > .5$, it has been shown that $MSE_{\infty}^{est}(1) < MSE_{\infty}^{est}(2)$. Consider now the interval $(.5, .504]$ of λ. In this case one can prove that the MSE increases monotonically with T from $T = 1$ to $T = 7$ and $MSE_{\infty}^{est}(8) \leq MSE_{\infty}^{est}(7)$. On the other hand, using Proposition 3, we know that the MSE decreases with T thereafter. Consequently, the optimal T can be computed by comparing $MSE_{\infty}^{est}(1)$ with $MSE_{\infty}^{est}(T \to \infty)$. But we already know that

[17] Again the right hand number of this interval, .446, is the (unique) value of λ at which $MSE_{\infty}^{est}(4) = MSE_{\infty}^{est}(5)$. One can now get the flavor as to how the succeeding intervals arise.

it can't be optimal to use $T \to \infty$. A more direct way of proving this is by comparing the two MSE, that is,

$$MSE^{est}_{\infty}(T \to \infty) < MSE^{est}_{\infty}(1)$$

$$\text{iff } \frac{1}{(1-\lambda)(1+\lambda)} < \frac{2}{(1+\lambda)}$$

$$\text{iff } \lambda < .5$$

This proves that when $\lambda \in (.5, .504]$, the optimal T is 1.

We then consider the interval $(.504, .521]$. In this case the MSE increases monotonically from $T = 1$ to $T = 8$ and, thereafter, decreases with T. The optimal T is, therefore, again 1. If we similarly consider the interval $(.521, .537]$, the MSE can be shown to increase monotonically with T from $T = 1$ to $T = 9$ and, thereafter, decrease with T so that the optimal T is again 1.

One can continue in this fashion and look at higher intervals of λ. Thus, when $\lambda = .88$, $T(\lambda) = 250$. Proposition 3 already tells us that the MSE decreases for all $T \geq 250$. On the other hand, it is possible to show that the MSE increases with T from $T = 1$ to $T = 250$. Consequently, it is still optimal to use $T = 1$.

Figure 1 a :

Plot of the (normalized) least squares residuals for 80 observations when $\lambda = .1$; T = 50; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped .

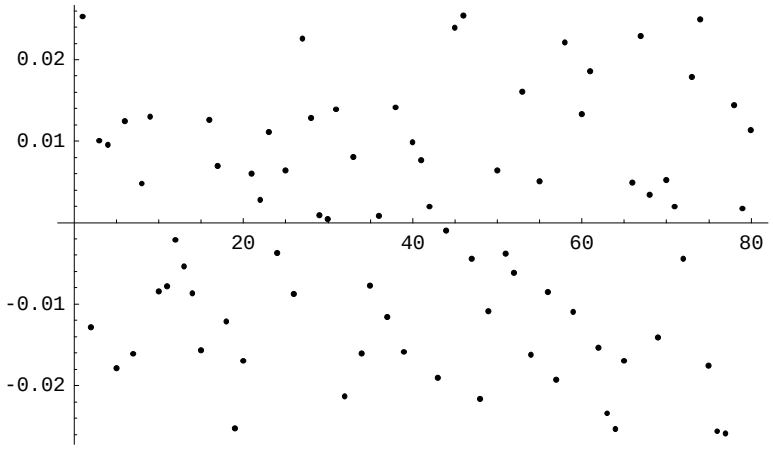

Figure 1 b :

Autocorrelation of the least squares errors for 80 observations when $\lambda = .1$; T = 50; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped . Straight lines indicate $\left\{+\frac{2}{\sqrt{M}}, -\frac{2}{\sqrt{M}}\right\}$ where M is the sample size . Only autocorrelation coefficients outside the straight lines would be considered significantly different from zero at the 5 % level .

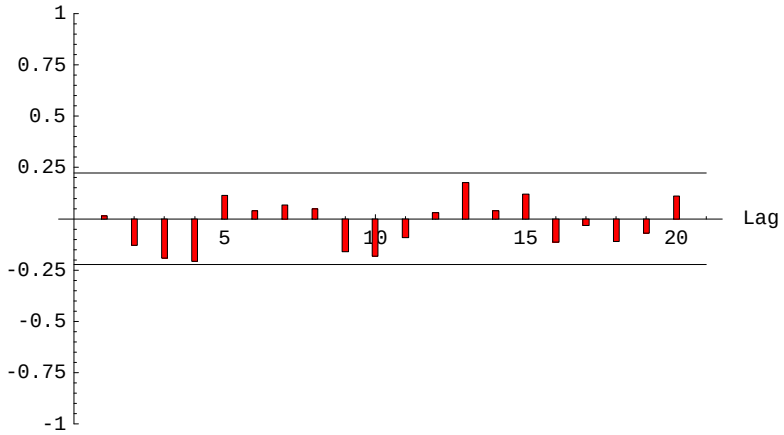

Figure 2 a :

Plot of the (normalized) least squares residuals for 35 observations when $\lambda = .6$; T = 1; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped .

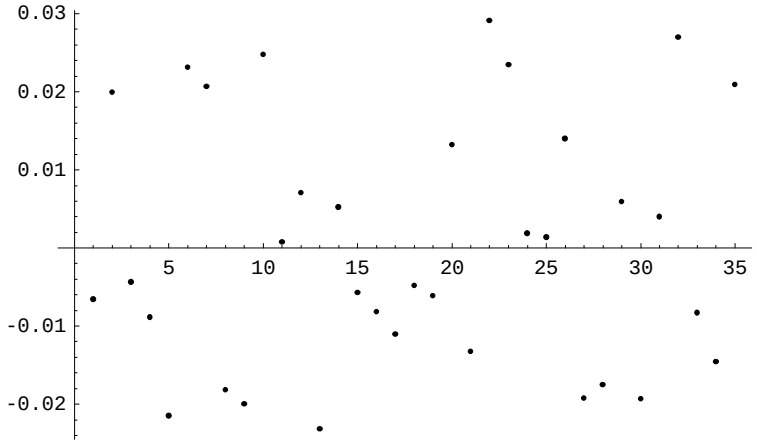

Figure 2 b :

Autocorrelation of the least squares errors for 35 observations when $\lambda = .6$; T = 1; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped . Straight lines indicate $\left\{+\frac{2}{\sqrt{M}}, -\frac{2}{\sqrt{M}}\right\}$ where M is the sample size . Only autocorrelation coefficients outside the straight lines would be considered significantly different from zero at the 5 % level .

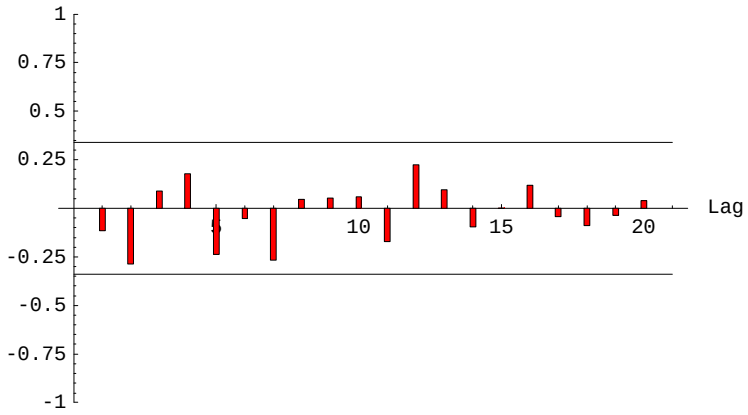

Figure 3 a :

Plot of the (normalized) least squares residuals for 35 observations when $\lambda = .8$; $T = 1$; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped .

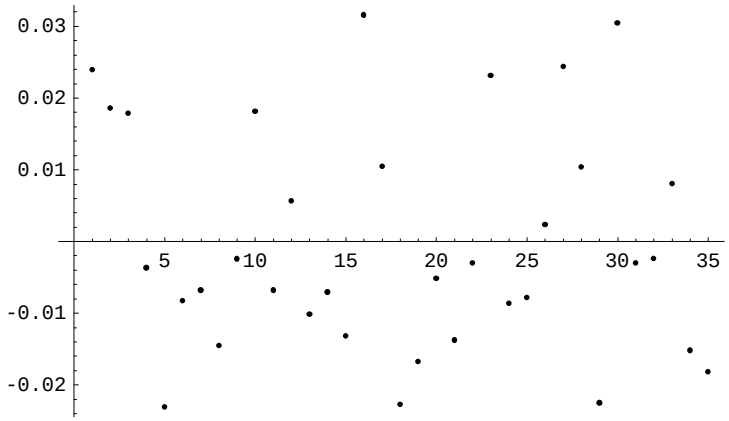

Figure 3 b :

Autocorrelation of the least squares errors for 35 observations when $\lambda = .8$; $T = 1$; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped . Straight lines indicate $\left\{+\frac{2}{\sqrt{M}}, -\frac{2}{\sqrt{M}}\right\}$ where M is the sample size . Only autocorrelation coefficients outside the straight lines would be considered significantly different from zero at the 5 % level .

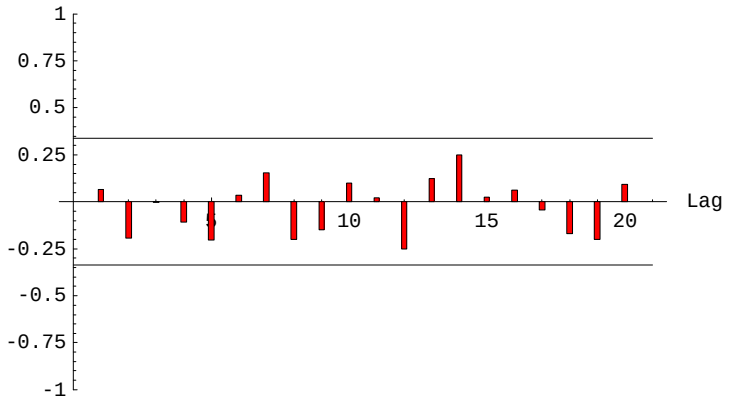

Figure 4 a :

Plot of the (normalized) least squares residuals for 100 observations when $\lambda = .6$; T = 50; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped .

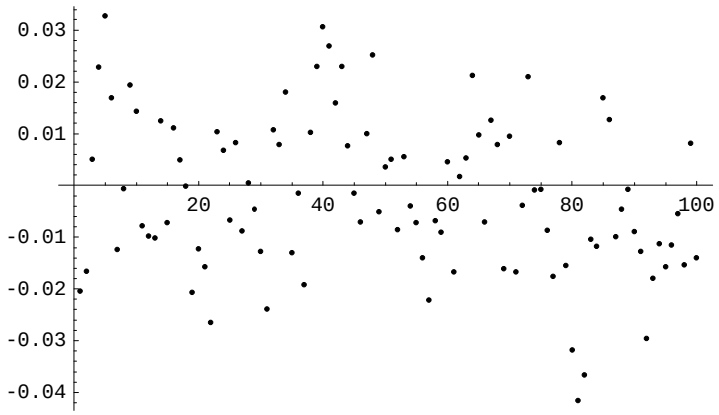

Figure 4 b :

Autocorrelation of the least squares errors for 100 observations when $\lambda = .6$; T = 50; noise is uniform with support $[-10^{-3}, 10^{-3}]$. The first 300 transients have been dropped . Straight lines indicate $\left\{+\frac{2}{\sqrt{M}}, -\frac{2}{\sqrt{M}}\right\}$ where M is the sample size . Only autocorrelation coefficients outside the straight lines would be considered significantly different from zero at the 5 % level .